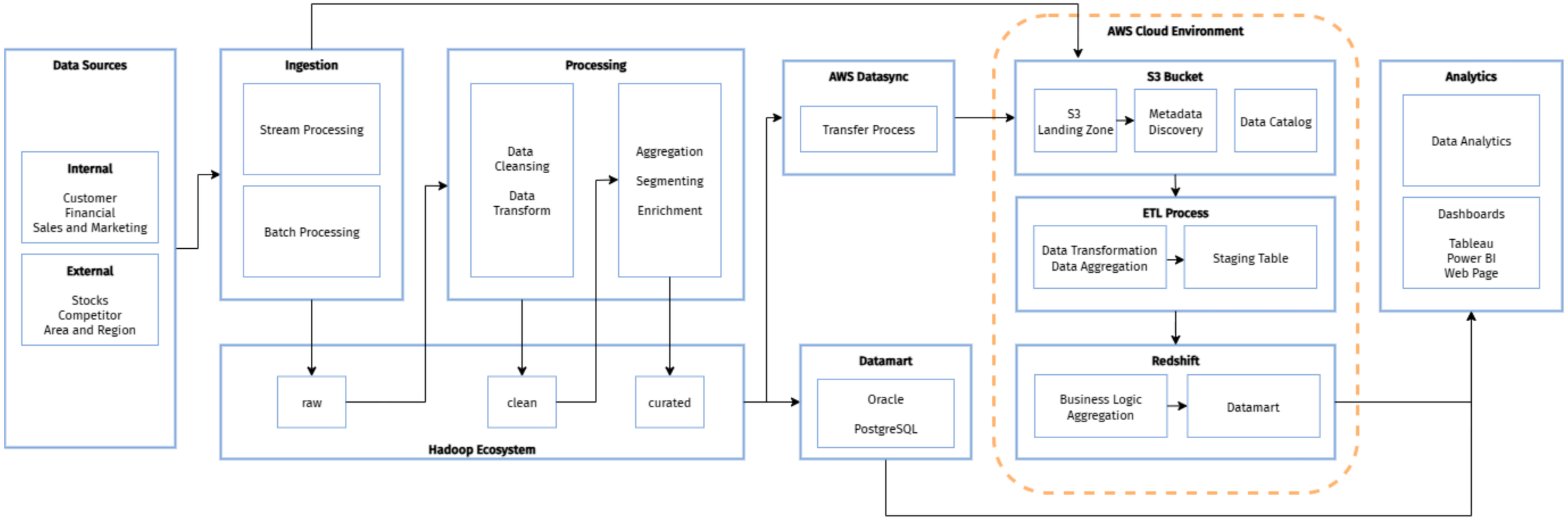
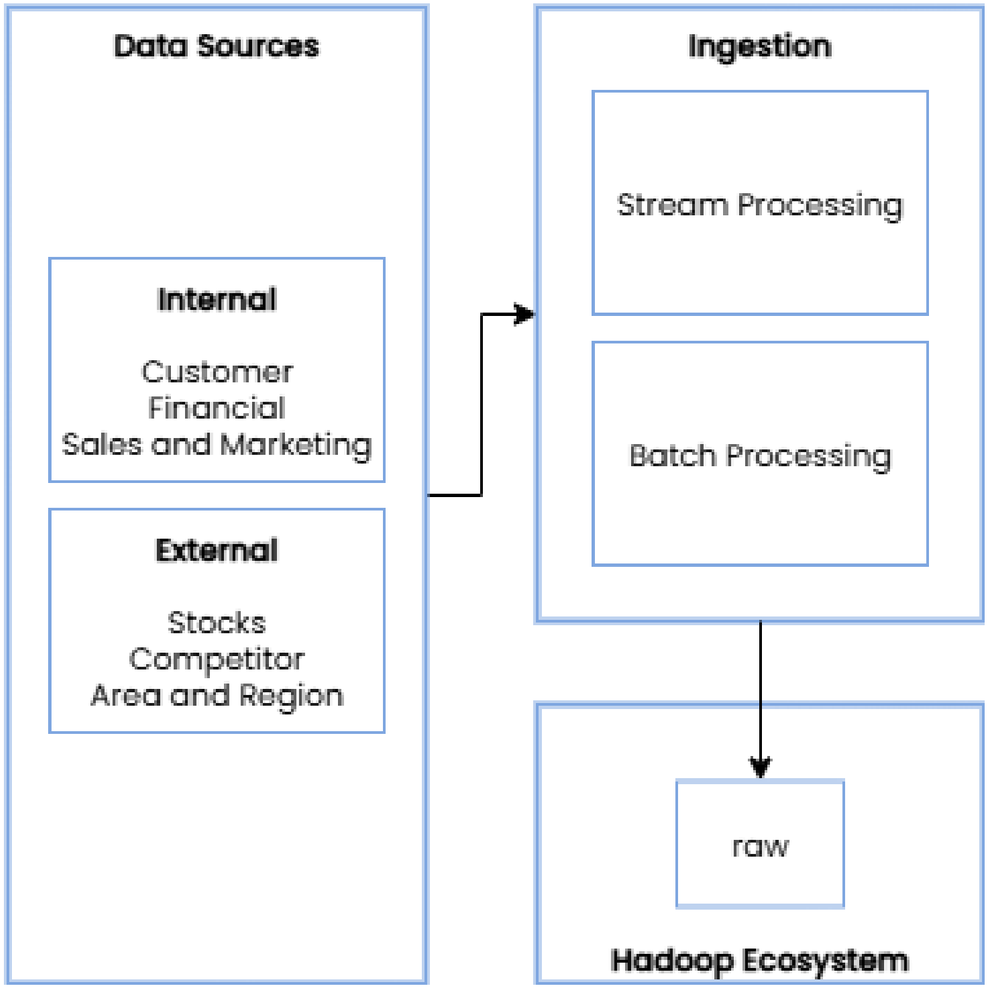
End-to-End Data Platform



Data Ingestion



Data acquisition: collect data from each source

- Internal → export from company's system (logs, ERP, CRM)
- External → collect from API's, public datasets

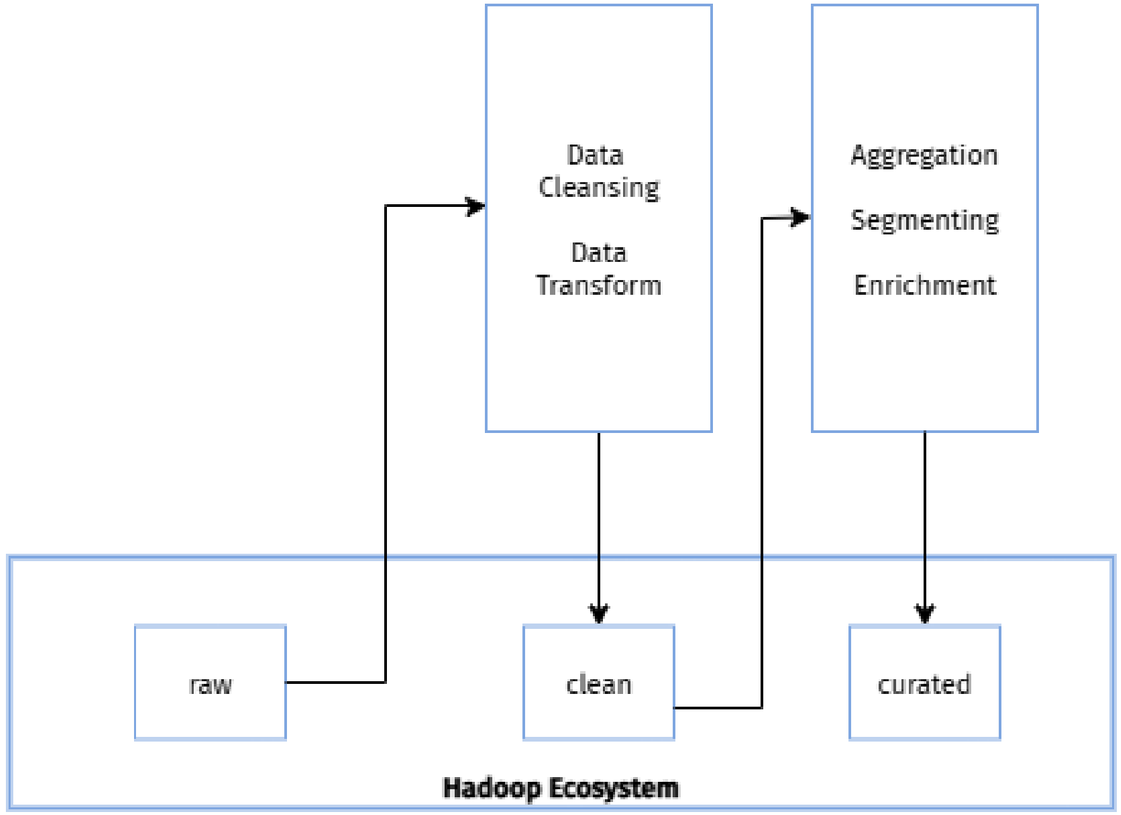
Convert to Hadoop formats

- Common format: Parquet, CSV, ORC

Send data to Hadoop

- Stream: Kafka, Flink
- Batch: Nifi

Data Processing



Data Cleansing

- Remove nulls
- Standardize formats
- Fix duplicates

Data Transformation

- Derive new columns
- Join with reference table
- Aggregate

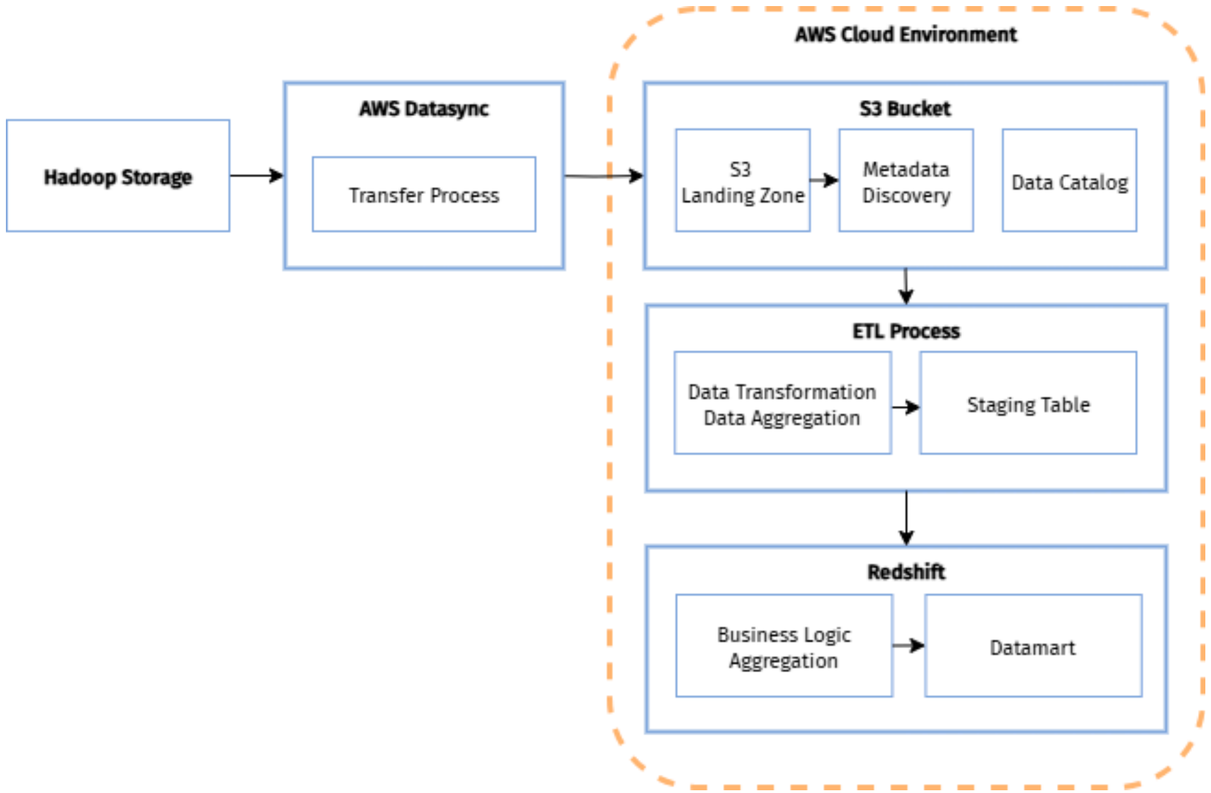
Data Modelling

- Using star schema

Tools

- Hive
- Spark

Cloud Process



Hadoop → S3 Bucket

- Deploy Datasync agent in on-premises env
- Configure HDFS as source and S3 bucket as destination
- Create schedule (optional)

S3 Bucket → Glue Crawler

- Pointing Glue Crawler to S3 landing path
- Crawler extract metadata and populates in Data Catalog

ETL Process

- Cleanse and transform data before load it into Redshift
- Use Glue Job to read from Data Catalog (S3 source)
- Put the data into table staging (optional)

Datamart

- Load data from staging into OLAP table using Stored Procedure